

## 프로야구 경기에서 안타의 독립성에 대한 연구<sup>†</sup>

김병수<sup>1</sup> · 박영욱<sup>2</sup> · 장나영<sup>3</sup>

<sup>123</sup>인제대학교 데이터정보학과/통계정보연구소

접수 2013년 10월 10일, 수정 2013년 11월 7일, 게재확정 2013년 11월 17일

### 요약

본 논문의 목적은 프로야구 경기에서 각 타석에서의 안타 여부가 이전 타석에서의 결과와 관련이 있는지를 살펴보는 것이다. 이를 위해 한국프로야구 8개 구단의 2011년 정규리그 자료를 이용하였다. 타석별로 안타에 대한 조건부 확률과 리프트를 구하여 본 결과 특정 타석의 타율이 이전 타석에서의 안타 여부에 따라 크게 달라지지 않는다는 것을 확인 할 수 있었다. 특정 타석과 이전 타석의 결과에 대한 독립성검정과 특정 타석의 결과와 이전 모든 타석에서의 무안타 여부와 독립성검정을 보면 거의 대부분의 검정에서 종속으로 볼 수 없다는 결과를 얻었다. 따라서 각 타석에서의 안타 여부는 이전 타석의 안타 여부와 관련성이 아주 적거나 서로 독립적인 관계에 있다고 결론 내릴 수 있다.

주요용어: 독립성, 리프트, 조건부 확률, 타율.

### 1. 서론

우리나라에서 프로스포츠는 1982년 프로야구를 시작으로 현재에는 축구, 골프, 복싱 등 다양한 종목에 걸쳐 시행되고 있다. 그중에서도 우리나라 프로스포츠의 시작을 알린 프로야구는 현재에도 남녀노소를 가리지 않고 많은 관중들의 관심으로 인기를 이어가고 있다. 그 인기는 여성 관객이 늘었다, 몇 백만 관중을 돌파했다는 등의 뉴스로 실감할 수 있다. 또한 휴대폰, TV, 인터넷, 라디오 등 경기장 관람이 아니라도 실시간으로 야구를 즐길 수 있는 수단이 늘어나면서 예전에 비해 야구를 쉽게 접할 수 있을 뿐 아니라 최근 세계 대회에서 좋은 성적을 거둔 것도 야구의 인기몰이에 한 몫을 한 것이라 볼 수 있다.

야구를 관람하는 사람들은 일차적으로는 응원하는 팀의 승패에 관심이 있겠지만, 더욱 열성적인 팬들은 세부적으로 들어가 야구를 관람하는 때 순간마다 다양한 작전을 생각하고 결과를 예상하는 등 단순히 야구를 관람 하는 것에서 그치지 않고 직간접적으로 경기에 참여하며 즐긴다. 또한 선수들의 타율, 타점, 득점 등 경기력에 관심을 가지고 분석하기에 이른다. 이러한 야구의 과학적인 특성 때문에 사람들은 야구를 기록의 경기, 확률의 경기라고 한다. 한국 프로야구에 관한 기존의 연구를 보면 프로야구 승패 추정에 관한 연구 (Cho 등, 2007; Choi와 Kim, 2011; Kim, 2001; Lee와 Bang, 2010)와 인공지능망을 이용하여 포스트시즌 진출 예측을 위한 연구도 수행되었다 (Chea 등, 2010). 또한 프로야구 선수들의 경기력과 연봉의 관계에 관한 연구 (Kim, 2002; Seung와 Kang, 2012; Sin, 2002), 출루 능력과 장타력이 득점 생산성에 미치는 영향에 관한 연구 (Kim, 2012; Gill와 Brajer, 1994), 프로야구 관중추이에

<sup>†</sup> 이 논문은 2009년도 인제연구장학재단 국외 (국내)연수지원에 의한 연구결과임.

<sup>1</sup> 교신저자 : (621-749) 경상남도 김해시 인제로 197, 인제대학교 데이터정보학과, 부교수.

E-mail: statkbs@inje.ac.kr

<sup>2</sup> (621-749) 경상남도 김해시 인제로 197, 인제대학교 데이터정보학과, 석사과정.

<sup>3</sup> (621-749) 경상남도 김해시 인제로 197, 인제대학교 데이터정보학과, 석사과정.

대한 연구 (Joен와 Nam, 2002) 등에서 보는 바와 같이 프로야구의 경기력과 선수계약, 마케팅, 관중추이 등에 대해 다양한 방법으로 연구들이 진행되고 있다.

타율, 득점, 타점, 도루 성공률, 실책 등 경기의 승패 여부에 영향을 미치는 요인들에 대한 연구들은 많지만, 본 논문에서는 각 타석에서의 안타의 독립성에 관심을 가지고 분석하고자 한다. 분석에 사용된 자료는 프로야구 8개 구단의 2011년 정규리그 자료를 이용하였다. 각 타석에서의 결과를 바탕으로 특정 타석의 안타 여부가 이전 타석의 결과와 관련성이 있는지 아니면 독립인지 알아보기 위하여 안타에 대한 조건부 확률이 조건이 없을 때의 타율과 같은지를 살펴보고, 리프트를 이용하여 특별한 특징이 있는지 확인하고, 독립성검정 (Fisher, 1925)을 이용하여 각 타석에서의 결과가 종속이라고 할 수 없음을 밝히고자 한다.

본 논문의 구성은 다음과 같다. 2절에서는 분석 자료에 대한 설명과 타율의 요약테이블에 대한 내용을 제시한다. 3절에서는 안타에 대한 조건부 확률과 리프트를 이용하여 타석별 타율과 비교하고, 독립성검정을 통하여 분석 결과를 도출한다. 마지막으로 4절에서는 본 논문의 결론을 제시한다.

## 2. 분석 자료

### 2.1. 분석 자료와 변수 설명

본 연구에서 사용한 원 자료는 KBO 홈페이지의 경기 결과 스코어보드에서 제공받았으며, 2011년 4월 2일부터 10월 6일까지 정규 시즌 야구 경기 데이터를 사용하였다. 8팀에 대하여 총 경기 수는 각 팀당 133경기로 총 532경기이다. 또한, 선수의 수는 총 209명에서 동명이인 (이병규) 2명과 이름은 등록되어 있으나 기록이 없는 선수 4명 (김경태, 김정남, 명재철, 신주영)을 제외한 203명의 자료를 사용하였다.

Table 2.1은 각 경기에서 한 선수의 경기 결과가 여러 칼럼에 걸쳐 나타나는 것을 행마다 각 타석의 결과가 나타나도록 정리한 자료의 일부이다. 또한 타석과 타수를 구별하는 변수를 생성하고, 타석 또는 타수의 결과 값을 0=아웃, 1=안타, 2=타수 제외로 나타내었다. 타수 제외는 4구 등으로 안타 여부와는 관계가 없는 것이며, 타수 제외를 제거한 것이 타수이며 타석 수 보다 적거나 같게 된다. 대타 등으로 선수가 바뀌는 경우에는 그 선수의 첫 타석이 1타석이 된다. Table 2.1에 사용된 변수는 Table 2.2와 같으며 PR은 결과, PA는 타석 순서, AB는 타수를 나타낸다.

**Table 2.1** Game record

Name	Date	Team	Inning	PR	PA	AB	Game
Garcia	18JUN2011	Hanwha	1	2	1	0	8
Garcia	18JUN2011	Hanwha	3	0	2	1	8
Garcia	18JUN2011	Hanwha	6	1	3	2	8
Garcia	19JUN2011	Hanwha	2	0	1	1	9
Garcia	19JUN2011	Hanwha	4	2	2	1	9
Garcia	19JUN2011	Hanwha	5	0	3	2	9
Garcia	19JUN2011	Hanwha	8	0	4	3	9

**Table 2.2** Variables description of the game record

Variable name	Variable description
Name	Player name
Date	Match date
Team	Name the team
Inning	Inning
PR (Plate Result)	The results of the plate with the category (0=out, 1=hit, 2=except at bats)
PA (Plate Appearance)	Plate Appearance
AB (At Bat)	At Bat
Game	Number of matches

## 2.2. 타석과 이닝별 타율

전체 선수에 대한 데이터로 타석별 타율과 이닝별 타율을 계산하였다. Table 2.3은 타석별로 총 타석 수, 총 안타 수, 타율로 구성되어 있으며, 타율 (hit rate)은 총 안타 수 (hit freq.)를 총 타석 수 (PA freq.)로 나눈 값이다. Table 2.4는 이닝별로 총 타석 수, 총 안타 수, 타율로 구성되어 있다.

Table 2.3을 보면 7타석까지의 기록이 있다. 타석 순이 커짐에 따라 타석 수와 안타 수가 적어짐을 볼 수 있다. 9번으로 케이스가 상대적으로 너무 적은 7타석을 제외한 나머지 타석의 타율을 살펴보면, 2타석에서의 타율이 0.241로 가장 높다.

**Table 2.3** Hit rate summary table by PA

PA order	PA freq.	hit freq.	hit rate
1	11882	2577	0.217
2	9942	2398	0.241
3	8804	2032	0.231
4	7031	1593	0.227
5	2402	535	0.223
6	227	47	0.207
7	9	4	0.444

Table 2.4는 9이닝까지의 정규이닝과 연장전으로 12이닝까지 간 경우를 모두 고려해 만든 이닝별 타율 요약표이며 10, 11, 12이닝의 타석수는 정규이닝인 경우에 비해 매우 적다. 그리고 9회 말에 이기고 있는 팀이 공격할 차례이거나 공격 중에 이기게 되는 점수를 얻게 되면 경기를 종료하기 때문에 9이닝이 다른 이닝에 비해 타석 수가 적은 것으로 보인다. 정규이닝에서의 타율을 보면, 8과 9이닝에서의 타율이 가장 낮고 3, 4, 5이닝에서의 타율이 대체로 높다.

**Table 2.4** Hit rate summary table by inning

Inning	PA freq.	hit freq.	hit rate
1	4632	1080	0.233
2	4434	998	0.225
3	4566	1109	0.243
4	4533	1097	0.242
5	4574	1112	0.243
6	4525	1001	0.221
7	4560	1057	0.232
8	4425	924	0.209
9	3381	686	0.203
10	364	75	0.206
11	205	32	0.156
12	98	15	0.153

## 3. 안타에 대한 조건부확률과 독립성검정

### 3.1. 조건부확률과 리프트

사건  $A$ 에 대한 사건  $B$ 의 조건부 확률은 어떤 사건  $B$ 가 일어났다는 조건하에서 사건  $A$ 가 일어날 확률을 의미하고 기호로는  $P(A|B)$ 로 표현한다. 리프트는 조건이 없을 때의 확률에 대한 조건이 있을 때의 확률의 비이며  $P(A|B)/P(A)$ 와 같다. 만일 사건  $A$ 가 일어날 확률  $P(A)$ 와 조건부 확률  $P(A|B)$ 가 같은 값을 가진다면  $A$ 와  $B$ 는 서로 독립이라고 할 수 있으며 리프트는 1이 된다. 여기서는  $H_i$ 를  $i$ 번째 타석에서 안타치는 사건으로,  $H_i^c$ 를  $i$ 번째 타석에서 안타를 못 치는 사건으로, 그리고  $B_j^c$ 를  $j$ 번째 이전

모든 타석에서 안타를 치지 못하는 사건으로 정의하겠다. 이후의 분석에서는 Table 2.3에서 보는 바와 같이 6타석 이후에는 타석수가 적기 때문에 5타석까지의 자료만 사용하였다.

**3.1.1. 조건부 확률  $P(H_j|H_i)$ 과 리프트**

Table 3.1은  $i$ 번째 타석에서 안타를 쳤다는 조건하에서  $j$ 번째 타석에서 안타를 칠 확률  $P(H_j|H_i)$ 과 리프트를 계산한 것이다. 조건부 확률에서 가장 높은 확률을 나타내는 것은 4타석에서 안타를 쳤을 때 5타석에서 안타를 칠 확률  $P(H_5|H_4)$ 로 0.257이며 그 다음으로 높은 확률을 나타내는 것은  $P(H_2|H_1)$ 이며 0.256으로 나타났다.

리프트는  $P(H_j|H_i)/P(H_j)$ 로 정의된다. Table 3.1에서 보는 바와 같이 리프트 값들은 대부분 1 근방의 값을 가지고 있어 타석 간에 안타여부가 독립적이라는 것을 시사하고 있다. 특히  $P(H_5|H_3)$ 과  $P(H_5)$ 는 0.223으로 같은 값으로 나타나 리프트가 1이며, 3타석에서 안타를 쳤을 때 5타석에서 안타를 칠 확률과 조건 없이 5타석에서 안타 칠 확률이 같다는 것을 알 수 있다.

**Table 3.1** Conditional probability  $P(H_j|H_i)$  and lift by PA

$i$	$j$	$j$ th PA freq.	hit freq. at $i$ th PA	hit freq. at $j$ th PA	num. of $H_j \cap H_i$	$P(H_j H_i)$	$P(H_j)$	Lift $P(H_j H_i)/P(H_j)$
1	2	9942	2277	2398	583	0.256	0.241	1.062
1	3	8804	2095	2032	487	0.232	0.231	1.004
1	4	7031	1771	1593	387	0.219	0.227	0.965
1	5	2402	735	535	154	0.210	0.223	0.942
2	3	8804	2243	2032	522	0.233	0.231	1.009
2	4	7031	1900	1593	423	0.223	0.227	0.982
2	5	2402	727	535	166	0.228	0.223	1.022
3	4	7031	1715	1593	373	0.217	0.227	0.956
3	5	2402	665	535	148	0.223	0.223	1.000
4	5	2402	639	535	164	0.257	0.223	1.152

**3.1.2. 조건부 확률  $P(H_j|H_i^c)$ 과 리프트**

Table 3.2는  $i$ 번째 타석에서 안타를 못 쳤다는 조건하에서  $j$ 번째 타석에서 안타를 칠 확률과 리프트를 계산한 것이다. 조건부 확률에서 가장 높은 확률을 나타내는 것은 1타석에서 안타를 못 쳤을 때, 2타석에서 안타를 칠 확률인  $P(H_2|H_1^c)$ 가 0.237로 가장 높은 확률을 나타냈다.

리프트는  $P(H_j|H_i^c)/P(H_j)$ 로 리프트 값이 1에 아주 근사한 값을 가지는 것을 확인 할 수 있다. 특히  $P(H_5|H_3^c)/P(H_4)$ 는 1.000으로 3타석에서 안타를 못 쳤을 때 5타석에서 안타를 칠 확률과 조건 없이 5타석에서 안타 칠 확률이 같다는 것을 알 수 있다.

**Table 3.2** Conditional probability  $P(H_j|H_i^c)$  and lift by PA

$i$	$j$	$j$ th PA freq.	no hit freq. at $i$ th PA	hit freq. at $j$ th PA	num. of $H_j \cap H_i^c$	$P(H_j H_i^c)$	$P(H_j)$	Lift $P(H_j H_i^c)/P(H_j)$
1	2	9942	7665	2398	1815	0.237	0.241	0.983
1	3	8804	6709	2032	1545	0.230	0.231	0.996
1	4	7031	5260	1593	1206	0.229	0.227	1.001
1	5	2402	1667	535	381	0.229	0.223	1.027
2	3	8804	6561	2032	1510	0.230	0.231	0.996
2	4	7031	5131	1593	1170	0.228	0.227	1.004
2	5	2402	1675	535	369	0.220	0.223	0.987
3	4	7031	5316	1593	1220	0.230	0.227	1.013
3	5	2402	1737	535	387	0.223	0.223	1.000
4	5	2402	1763	535	371	0.210	0.223	0.942

**3.1.3. 조건부 확률  $P(H_j|B_j^c)$ 와 리프트**

Table 3.3은  $j$ 번째 이전 모든 타석에서 안타가 없다는 조건하에서  $j$ 번째 타석에서 안타를 칠 확률  $P(H_j|B_j^c)$ 과 리프트를 계산한 것이다. 특정 타석 이전까지 안타를 치지 못한 결과가 계속해서 이어지는지 또는 반대로 안타를 칠 확률이 커지는지를 해당 타석의 타율과 비교하여 확인 할 수 있다. Table 3.3에 나타난 값을 보면 특정 타석 이전 모든 타석에서 안타를 치지 못했다는 조건하에서 안타를 칠 확률이 조건이 없을 때의 타석별 타율과 크게 차이가 나지 않는다는 것을 알 수 있다. 리프트 값들 또한 모두 1 근방의 값을 가지는 것을 확인 할 수 있다.

**Table 3.3** Conditional probability  $P(H_j|B_j^c)$  and lift by PA

$j$	$j$ th PA freq.	no hit freq. until $(j-1)$ th PA	num. of $H_j \cap B_j^c$	$P(H_j B_j^c)$	$P(H_j)$	Lift $P(H_j B_j^c)/P(H_j)$
2	9942	7665	1815	0.237	0.241	0.982
3	8804	5020	1154	0.230	0.231	0.996
4	7031	2919	688	0.236	0.227	1.040
5	2402	599	130	0.217	0.223	0.974

**3.2. 독립성 검정**

한 변수의 각 수준에 따라 다른 변수의 분포가 동일할 때 두 변수는 통계적으로 독립 (statistically independent)이라 한다. 일반적으로 두 변수가 서로 연관을 맺고 있는지 관심을 갖게 될 때 변수들 간의 연관성 검증에 사용되는 기법이 독립성 검정이다. 여기서는 각 변수들이 범주형이므로 독립성 검정으로 2차원 교차표와 피어슨검정통계량을 이용한  $\chi^2$ -검정 (chi-square test)을 실시하였고 유의수준은 0.05로 하였다.  $l \times m$  2차원 분할표의 독립성검정에서 피어슨통계량은 다음과 같이 나타낼 수 있다.

$$\chi^2 = \sum_{j=1}^m \sum_{i=1}^l \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

3.1.1절과 3.1.2절에서  $P(H_j|H_i)$ 와  $P(H_j|H_i^c)$ 의 값이  $P(H_j)$ 와 큰 차이가 없고 리프트가 1 근처의 값을 가진다는 것을 보였다. Table 3.4는 이전 타석의 안타 여부 (1=안타, 0=4구, 아웃 등)와 이후 타석의 안타 여부가 독립인지를 살펴 본 독립성 검정의 결과이다. 4타석과 5타석에 대한 독립성 검정에서 p값이 0.016으로 독립으로 보기 어려우나 이를 제외한 나머지 타석 간에 대한 독립성 검정 결과 모두 독립이 아니라고 볼 수 없다.

**Table 3.4** Chi-square test between previous PA and subsequent PA

previous	subsequent	PA freq.	$\chi^2$	p-value
1	2	9942	3.554	0.059
1	3	8804	0.042	0.837
1	4	7031	0.875	0.350
1	5	2402	1.067	0.302
2	3	8804	0.063	0.803
2	4	7031	0.230	0.631
2	5	2402	0.189	0.664
3	4	7031	1.066	0.302
3	5	2402	0.000	0.990
4	5	2402	5.786	0.016

3.1.3절에서  $P(H_j|B_j^c)$ 의 값이  $P(H_j)$ 와 큰 차이가 없고 리프트가 1 근처의 값을 가진다는 것을 보였다. Table 3.5는 특정 타석 이전까지의 안타 여부 (0=무안타, 1=한번이라도 안타)와 특정 타석의 안타 여부가 독립인지를 살펴 본 독립성 검정의 결과이다. 모든 경우 유의확률 값이 유의수준 0.05보다 작지

않기 때문에 특정 타석 이전까지의 무안타 여부와 특정 타석의 안타 여부는 관련이 없거나 아주 작다고 볼 수 있다.

**Table 3.5** Chi-square test between cumulative PA and subsequent PA

previous	subsequent	PA freq.	$\chi^2$	p-value
1	2	9942	3.554	0.060
1 ~2	3	8804	0.056	0.813
1 ~3	4	7031	2.374	0.123
1 ~4	5	2402	0.150	0.699

#### 4. 결론 및 향후과제

본 논문의 목적은 프로야구 경기에서 각 타석에서의 안타 여부가 독립적인지 살펴보는 것이다. 이를 위해 한국프로야구 8개 구단의 2011년 정규리그 자료를 이용하였으며 선수들의 각 타석에서의 결과를 나타내는 변수 등을 사용하였다. 각 선수들의 타석별 타율과 해당 경기에서 이전 타석의 결과를 이용하여 각 타석에서의 안타 여부가 이전 타석에서의 안타 여부와 관련이 있는지 아니면 각 타석에서의 안타 여부는 서로 독립인지 밝히기 위하여 독립성 검정을 실시하였다.

검정에 앞서 타석별로 안타에 대한 조건부 확률과 리프트를 구하여 이전 타석의 안타 여부가 특정 타석의 안타 여부에 영향을 주는지 확인하였다. 안타에 대한 조건부 확률의 경우 조건이 없을 때의 타율과 크게 차이하지 않았다. 또한 리프트 값 역시 대부분 1 근처의 값을 가지는 것을 확인하였다. 즉, 특정 타석의 안타에 대한 조건부 확률과 특정 타석의 타율이 큰 차이가 없다는 것을 확인하였으며 타석 간에 안타여부가 독립적이라는 것을 시사하고 있다.

안타에 대한 조건부 확률과 리프트의 결과를 바탕으로 특정 타석에서의 안타 여부는 이전 타석의 결과와 독립인지를 알기 위하여 각 타석에서의 안타 여부에 대해 독립성 검정을 실시하였다. 4타석과 5타석에 대한 결과를 제외하고 나머지 타석 간 독립성 검정 결과에서는 유의확률이 유의수준보다 크기 때문에 독립이 아니라고 볼 수 없었다. 또한 특정 타석 이전까지 모두 무안타인 것과 특정 타석의 안타 여부에 대해 독립성 검정 결과도 마찬가지로 독립이 아니라고 볼 수 없었다.

결론적으로 특정 타석의 안타 여부와 이전 타석의 안타 여부는 독립이거나 관련성이 아주 작다고 볼 수 있다. 본 논문에서는 제시하지 않았지만 경기 출전수가 많은 40명의 선수들에 대해서 그리고 타석이 아닌 타수를 기준으로 한 경우에도 전체 선수를 이용하여 타석에 대해 분석한 경우와 흡사한 결과를 얻었다. 따라서 각 타석 또는 타수에서의 안타 여부는 서로 독립이라고 보아도 무방할 것으로 사료된다.

본 논문에서는 이전 타석의 안타 여부만 이용하여 타석의 독립성에 대해 연구하였지만 향후 과제로 포지션, 이닝, 타석, 타수, 지난 경기의 타율 등 여러 가지 변수를 적용하여 특정 선수의 특정 타석에서의 안타 가능성을 예측하는 모형에 대한 연구가 필요하겠다.

#### References

- Chea, J. S., Cho, E. H. and Eom, H. J. (2010). Comparisons of the outcomes of statistical models applied to the prediction of post-season entry in Korean professional baseball. *The Korean Journal of Measurement and Evaluation in Physical Education and Sport Science*, **12**, 33-48.
- Cho, Y. S., Cho, Y. J. and Shin, S. K. (2007). A study on winning and losing in Korean professional baseball league. *Journal of the Korean Data Analysis Society*, **1**, 501-510.
- Choi, Y. G. and Kim, H. M. (2011). A statistical study on Korean baseball league games. *The Korean Journal of Applied Statistics*, **24**, 915-930.
- Fisher, R. A. (1925). *Statistical methods for research workers*, Oliver and Boyd, Edinburgh.

- Gill, A. M. and Brajer, V. (1994). Baseball stars and baseball cards: A new look at monopsony in major league baseball. *Social Science Quarterly*, **75**, 195.
- Jeon, Y. B. and Nam, J. W. (2002). The study on the attendance changes and profit structures in Korean professional baseball. *Journal of Sport and Leisure Studies*, **18**, 393-407.
- Kim, C. Y. (2001). A win-loss predicting model by analyzing professional baseball game. *Journal of Sport and Leisure Studies*, **16**, 807-819.
- Kim, E. S. (2002). The relationship of game performance and annual salary for Korean professional baseball pitchers. *Korean Society for the Sociology of Sport*, **15**, 95-104.
- Kim, H. J. (2012). Effects of on-base and slugging ability on run productivity in Korean professional baseball. *Journal of the Korean Data & Information Science Society*, **23**, 1165-1174.
- Lee, J. T. and Bang, S. Y. (2010). Forecasting attendance in the Korean professional baseball league using garch models. *Journal of the Korean Data & Information Science Society*, **21**, 1041-1049.
- Seung, H. B. and Kang, K. H. (2012). A study on relationship between the performance of professional baseball players and annual salary. *Journal of the Korean Data & Information Science Society*, **23**, 285-298.
- Sin, M. S. (2002). A comparison on methods for estimating player's annual salaries in Korean professional soccer league and Korean professional baseball league. *Korean Society of Sport Management*, **7**, 141-155.

## Study for independence of hits in professional baseball games<sup>†</sup>

Byungsoo Kim<sup>1</sup> · Youngwook Park<sup>2</sup> · Nayoung Jang<sup>3</sup>

<sup>123</sup>Department of Data information / Institute of Statistical Information, Inje University

Received 10 October 2013, revised 7 November 2013, accepted 17 November 2013

### Abstract

In this paper, we would like to test whether the hit at a particular bat has a dependency with the hitting results at the previous bats in professional baseball games. For this purpose, we used the 2011 Korean Baseball League data. We find out that the hitting percentage at a particular bat has no dependency with the hit at the previous bat, after reviewing the conditional probability of hit at each bat and the lift. From the independence test of hits at consecutive bats, and hit at a particular bat with no hits at previous bats, we can conclude that hits at particular bats are not dependent on the hits at previous bats in most cases. Hence, we can safely conclude that a hit at a particular bat is statistically independent from the hits at the previous bats.

*Keywords:* Conditional probability, independence, hit rate, lift.

---

<sup>†</sup> This work was supported by the Inje Research and Scholarship Foundation in 2009.

<sup>1</sup> Corresponding author: Associate professor, Department of Data Science, Inje University, Gyeongnam 621-749, Korea. E-mail : statkbs@inje.ac.kr

<sup>2</sup> Graduate student, Department of Data Science, Inje University, Gyeongnam 621-749, Korea.

<sup>3</sup> Graduate student, Department of Data Science, Inje University, Gyeongnam 621-749, Korea.